# Web Scraping Final Project Description - Filmweb

## 1. Group:

Edyta Pszczółkowska 435022
Wiktor Głuszek 387182

## 2. Short description:

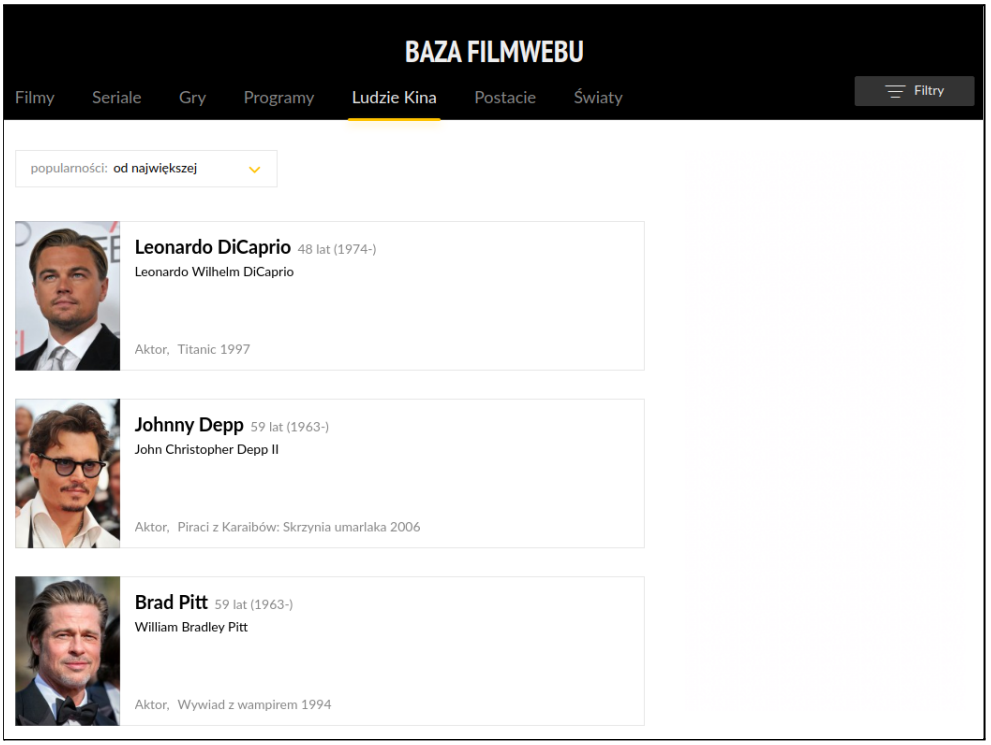
The main aim of this project was to scrape information from filmweb about the most popular actors and directors (mainly actors). The data scraped include variables such as:

- Person's name,
- Birthday,
- Filmweb's score (grade),
- Number of opinions about person
- Films from which actor / director is known for
- Awards
- Nominations
- Best roles
- Ranking
- Newest productions

The Filmweb website is static, therefore we used 3 different scrapers: Beautifulsoup, Selenium and Scrapy.

## 3. Short description of your scraper mechanics:

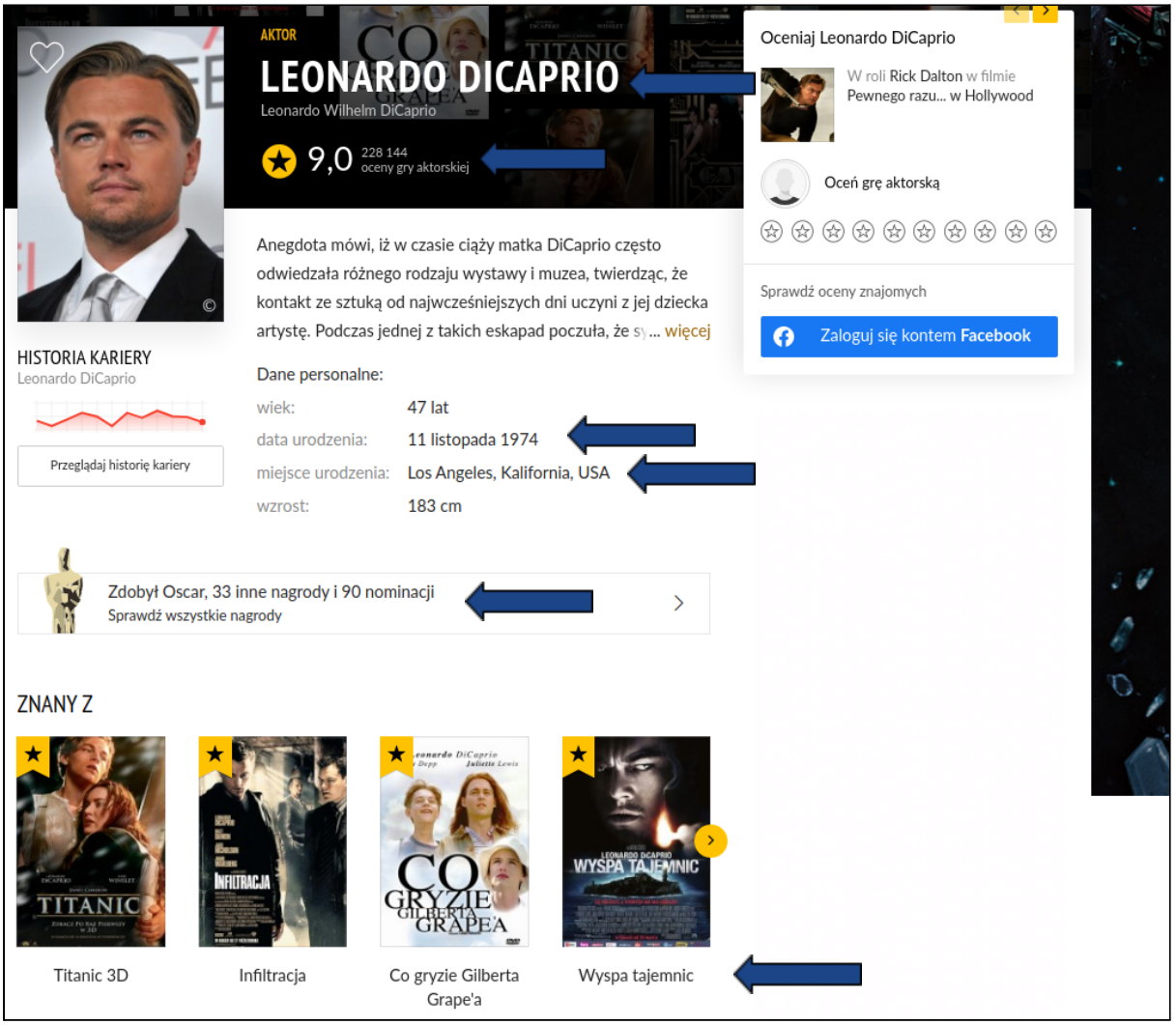
Firstly, scrapers enter the main list of actors, from which links to subpages are extracted.




There are 10 actors per page, scrapers alter the main link (....page="2" etc.) in order to access other pages.

All links are saved in a list, then scrapers enter each specific subpage one by one.

All variables are scraped from subpages. (examples highlighted below)




**4. Short technical description of the output:**

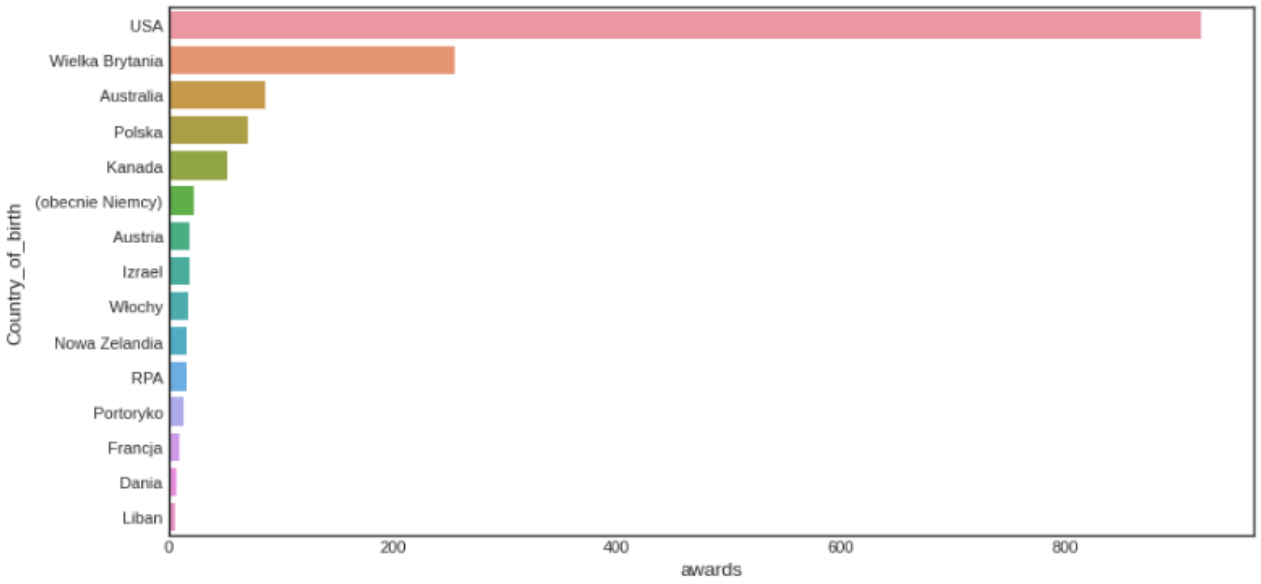
The output is in the form of a table which consists of all scraped variables. One column one variable. Before displaying the final table some cleaning is being done (this step may vary depending on scraper). For example, replacing commas with dots in “grade”, removing space in no_opinions and splitting ‘awards_str’ (long string with combined info) into awards and nominations. Example of output below (part).

| | name | birthday | birth_location | grade | no_opinions | awards_str | known_for | best_roles | newest_productions | ranking | awards | nominations | Countr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Leonardo DiCaprio | 11 listopada 1974 | Los Angeles, Kalifornia, USA | 9.0 | 228106 | Zdobył Oscar, 33 inne nagrody i 90 nominacji | Titanic 3D, Infiltracja, Co gryzie Gilberta Gr... | Arnie Grape, Jordan Belfort, Jim Carroll, Tedd... | Roosevelt, Killers of the Flower Moon, Jim Jon... | 3 | 34 | 90 | |
| 1 | Johnny Depp | 9 czerwca 1963 | Owensboro, Kentucky, USA | 8.8 | 214508 | Zdobył Złoty Glob, 23 inne nagrody i 73 nominacje | Piraci z Karaibów: Skrzynia umarlaka, Piraci z... | Raoul Duke, Sweeney Todd, Sam, James Barrie, | In the Hand of Dante, Minamata, Czekając na ba... | 18 | 24 | 73 | |
| 2 | Brad Pitt | 18 grudnia 1963 | Shawnee, Oklahoma, USA | 8.7 | 162566 | Zdobył 2 nagrody Oscar, 34 inne nagrody i 70 n... | Wywiad z wampirem, Pan i pani Smith, Zniewolon... | Tyler Durden, Jeffrey Goines, Mickey O'Neil, E... | Bullet Train, Zaginione miasto, Babylon, Fast ... | 29 | 36 | 70 | |
| 3 | Robert De Niro | 17 sierpnia 1943 | Nowy Jork, Nowy Jork, USA | 8.7 | 150708 | Zdobył 2 nagrody Oscar, 20 innych nagród i 56 ... | Taksówkarz, Wściekły Byk, Chłopcy z ferajny, P... | Leonard Lowe, Travis Bickle, Jake La Motta, Mi... | Killers of the Flower Moon, Untitled David O. ... | 36 | 22 | 56 | |
| 4 | Samuel L. Jackson | 21 grudnia 1948 | Waszyngton, Dystrykt Kolumbii, USA | 8.4 | 143246 | Zdobył BAFTA, 6 innych nagród i 31 nominacji | Pulp Fiction, Django, Nienawistna ósemka, Jack... | Major Marquis Warren, Trener Ken Carter, Dariu... | The Marvels, Argylle, The Blob, Secret Invasio... | 114 | 7 | 31 | |

## 5. Data analysis

As shown on the chart below, the most number of awards is won by actors from (born in) the USA, secondly from Great Britain. Poland is in 4th place, however Filmweb is a Polish website, therefore scraped data may consist of more Polish actors due to their popularity in Poland (data was scraped starting from most popular to least popular actors).

**Chart 1. Sum of awards by country of birth**



Based on the data presented on the chart 2, one can conclude that Joaquin Phoenix is the most popular actor on filmweb with the highest grade, followed by Leonardo Di Caprio and Jack Nicholson. The “step” pattern is caused by the presentation of ‘grade’ on the webpage. ‘Grade’ has only one decimal place, therefore there are many actors who have the same score.

**Chart 2. Grade by actor (top 70 observations)**

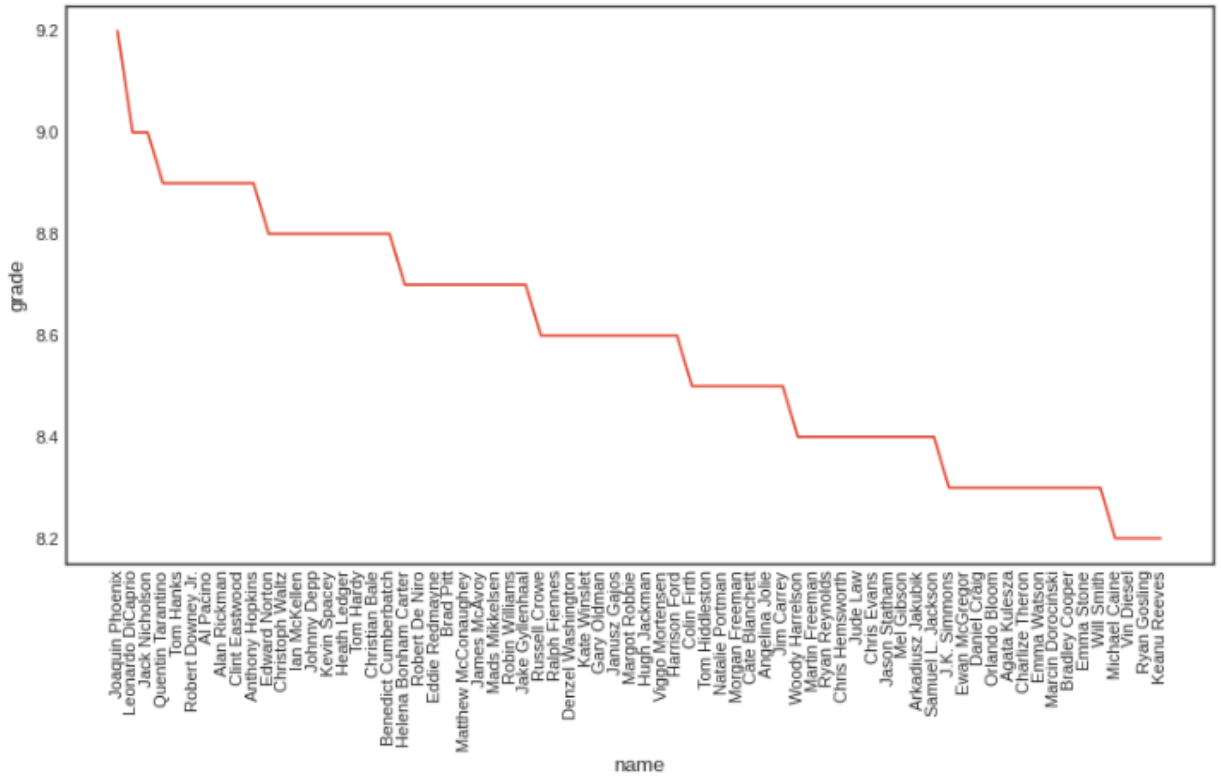
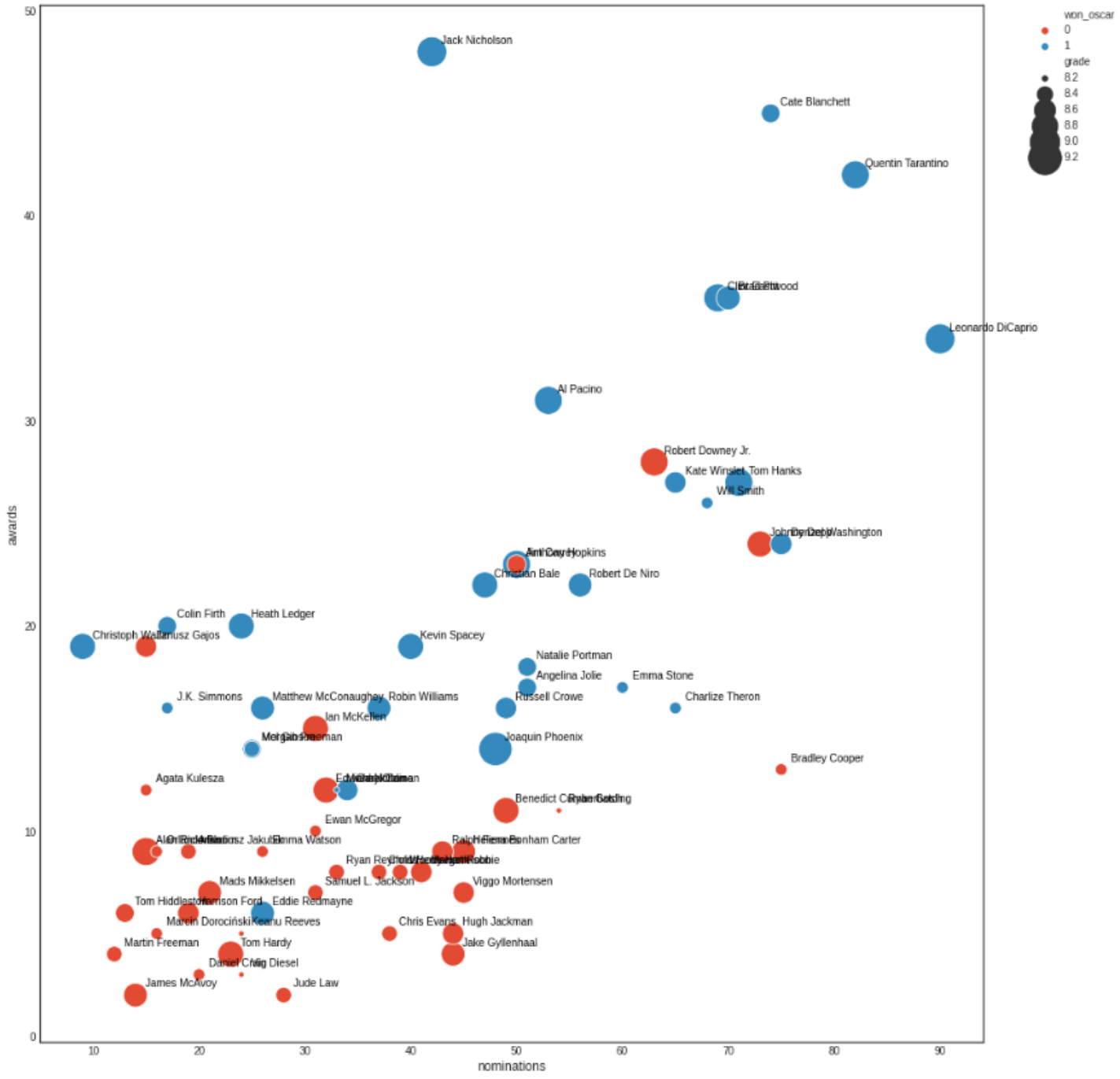



Chart 3 (below) presents different actors with nominations on x axis, awards on y axis, bubble size indicates grade and color if a given actor won an Oscar award.

Based on this visualization, one can conclude that:

- Rober Downey Jr. has many awards and nominations, however he didn’t get an Oscar.
- Actors with less than 50 nominations and less than 12 awards usually do not win oscars.
- Cate Blanchett has a lot of awards and nomination, however she is not so popular based on ‘grade’
- Christoph Waltz if nominated, wins an award in ~ 2/3 of cases

**Chart 3. Awards, nominations and grade by actor (top 70 observations)**




**6. Detailed description which group participant wrote which part of the project.**

Edyta - Scrapy scraper and Beutfiulsoup scraper

Wiktor - Selenium Scraper and PDF Report and data analysis